SCIENTIFIC AMERICAN.

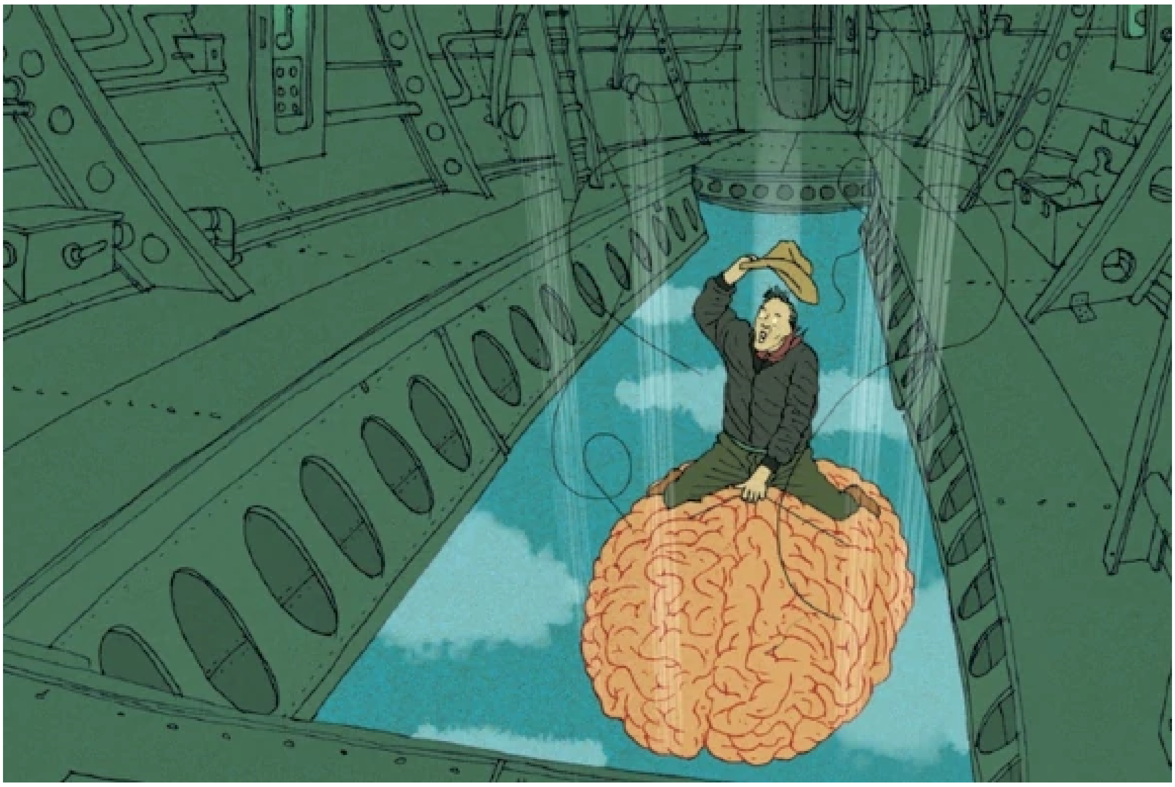
SUBSCRIBE

ENGINEERING

Artificial Intelligence Is Not a Threat—Yet

Artificial intelligence as existential threat

By Michael Shermer | Scientific American March 2017 Issue



Credit: Izhar Cohen

VISIT…

LANZAROTE

Caliente.COM

In 2014 SpaceX CEO Elon Musk tweeted: “Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes.” That same year University of Cambridge cosmologist Stephen Hawking told the BBC: “The development of full artificial intelligence could spell the end of the human race.” Microsoft co-founder Bill Gates also cautioned: “I am in the camp that is concerned about super intelligence.”

How the AI apocalypse might unfold was outlined by computer scientist Eliezer Yudkowsky in a paper in the 2008 book *Global Catastrophic Risks:* “How likely is it that AI will cross the entire vast gap from amoeba to village idiot, and then stop at the level of human genius?” His answer: “It would be physically possible to build a brain that computed a million times as fast as a human brain.... If a human mind were thus accelerated, a subjective year of thinking would be accomplished for every 31 physical seconds in the outside world, and a millennium would fly by in eight-and-a-half hours.” Yudkowsky thinks that if we don't get on top of this now it will be too late: “The AI runs on a different timescale than you do; by the time your neurons finish thinking the words ‘I should do something’ you have already lost.”

The paradigmatic example is University of Oxford philosopher Nick Bostrom's thought experiment of the so-called paperclip maximizer presented in his *Superintelligence* book: An AI is designed to make paperclips, and after running through its initial supply of raw materials, it utilizes any available atoms that happen to be within its reach, including humans. As he described in a 2003 paper, from there it “starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities.” Before long, the entire universe is made up of paperclips and paperclip makers.

I'm skeptical. First, all such doomsday scenarios involve a long sequence of if-then contingencies, a failure of which at any point would negate the apocalypse. University of West England Bristol professor of electrical engineering Alan Winfield put it this way in a 2014 article: “*If* we succeed in building human equivalent AI and *if* that AI acquires a full understanding of how it works, and *if* it then succeeds in improving itself to produce super-intelligent AI, and *if* that super-AI, accidentally or maliciously, starts to consume resources, and *if* we fail to pull the plug, then, yes, we may well have a problem. The risk, while not impossible, is improbable.”

Second, the development of AI has been much slower than predicted, allowing time to build in checks at each stage. As Google executive chairman Eric Schmidt said in response to Musk and Hawking: “Don't you think humans would notice this happening? And don't you think humans would then go about turning these computers off?” Google's own DeepMind has developed the concept of an AI off switch, playfully described as a “big red button” to be pushed in the event of an attempted AI takeover. As Baidu vice president Andrew Ng put it (in a jab at Musk), it would be “like worrying about overpopulation on Mars when we have not even set foot on the planet yet.”

Third, AI doomsday scenarios are often predicated on a false analogy between *natural intelligence* and *artificial intelligence*. As Harvard University experimental psychologist Steven Pinker elucidated in his answer to the 2015 Edge.org Annual Question “What Do You Think about Machines That Think?”: “AI dystopias project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or taking over the world.” It is equally possible, Pinker suggests, that “artificial intelligence will naturally develop along female lines: fully capable of solving problems, but with no desire to annihilate innocents or dominate the civilization.”

Fourth, the implication that computers will “want” to do something (like convert the world into paperclips) means AI has emotions, but as science writer Michael Chorost notes, “the minute an A.I. *wants* anything, it will live in a universe with rewards and punishments—including punishments from us for behaving badly.”

Given the zero percent historical success rate of apocalyptic predictions, coupled with the incrementally gradual development of AI over the decades, we have plenty of time to build in fail-safe systems to prevent any such AI apocalypse.

This article was originally published with the title "Apocalypse AI"